Data 583 Life Expectancy (WHO)

Justin Chan, Kenny Tong, Viji Rajagopalan

7 Mar, 2023

Data 583 Life Expectancy - Exploratory Data Analysis

Life Expectancy

Introduction

We selected “this” dataset and our goal is to understand LE. in this document, we will take a detailed look at different aspects about variables, summary and apply statistical techniques to understand the underlying data more. Also, we will come up with next steps of how to fine tune data further and improve modeling <>. (Some 4 to 5 lines in total)

Data Exploration

Original Dataset Summary & Initial Data Screening

Purpose

- Let’s take a snapshot of the original dataset and have a rough idea of its record

Procedure Take a look at the dataset summary.

Then look at the dataset dimension.

Here is another overview :

```
## ’data.frame':    2938 obs. of  22 variables:
##  $ Country                        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
##  $ Status                         : chr  "Developing" "Developing" "Developing" "Developing" ...
##  $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
##  $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
##  $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
##  $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
##  $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
##  $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
##  $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
##  $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
##  $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
##  $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
##  $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
##  $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
##  $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
##  $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
##  $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
##  $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ...
##  $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

Conclusion/Key Findings :

- The records range is from Year 2000 to 2015
- Columns with NA : Life Expectancy, Adult Mortality, Alcohol, Hep B, BMI, Polio, Total exp, Dip, GDP, Population, thinness..1.19, thinness.5.9, Income.composition.of.resources, Schooling
- 'Status' Column is of the "character" data type, with values "Developing" and "Developed". We will introduce a new column 'Status.val' to be the factor value of 'Status' for better analysis..
- 'Percentage Expenditure' has a mean value of 738.2512955. Spending on health is more than the GDP per capita (!?). Look into the column definition : Expenditure on health as a percentage of Gross Domestic Product per capita(%). The data of such magnitude simply does not quite make sense. Cross check with other references (e.g. the World Bank https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS). OK, let's conclude that we have hesitation about the reliability/interpretation of the value of this column, and this column would be dropped for the rest of this analysis.

Null Value Analysis and Handling

Purpose

- Investigate the and determine how to handle the null value in the data set. Missing values could have a large affect to the overall quality of the static models and machine learning models and need to be clean before using it in our training model.

Procedure Lets investigate how many missing values within our features.

A brief check indicates that there are total of 2563 missing value within our dataset, we could visualize the missing data to identify patterns or cluster of missing values within our data to determine the cause of the missing data and whether it is random or systematic and to highlight potential biases that may exist in our data set. Visualizing the missing value also allow to understand the extend of the missing data and determine appropriate strategies for imputing missing value, since different imputation methods could be more appropriate depending on the pattern of the missing data.

There are 2938 no. of rows in the dataset. According to our Visualization, there seems to be a correlation in the appearance in missing data in our original data's feature "population", "gdp" , "income.composition.of.resources" and "schooling". We would deal with this correlation in missing data by removing the the record that have missing value in all of the listed variables.

For the other values, we would set the na to the respective column mean for the subsequent analysis.

Conclusion/Key Findings :

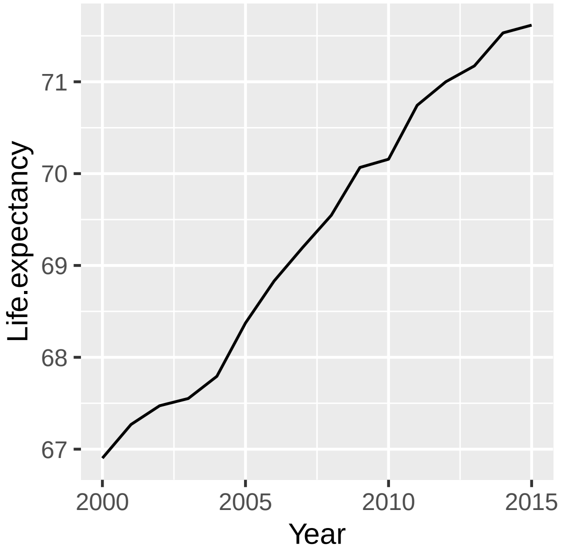
- na values have been analyzed
- Data imputation have been performed as far as possible in order to prepare for the subsequent data analysis.

Statistical Analysis

Purpose : Do some visualization to explore and identify the general data pattern, trends and clusters, etc

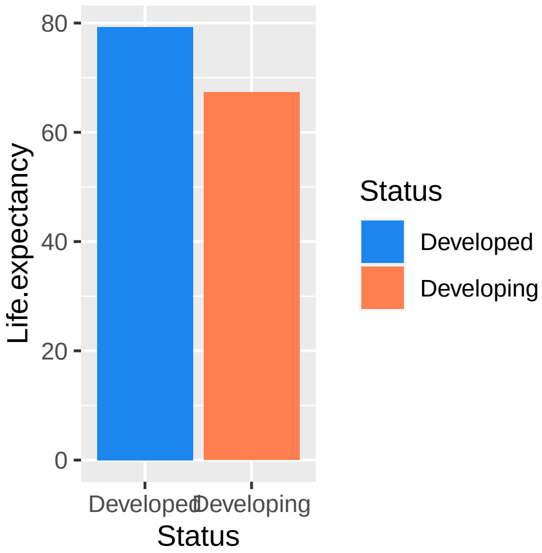
General Life Expectancy

As we are interested in Life Expectancy as our response variable, we first start looking at the distribution of the variable and general trend.



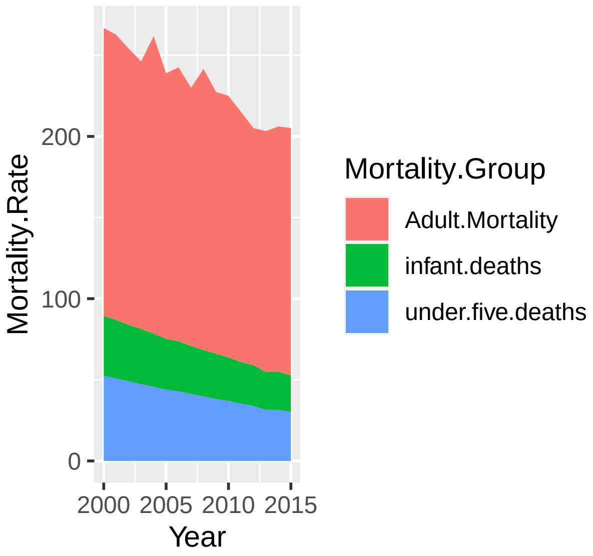
Findings :

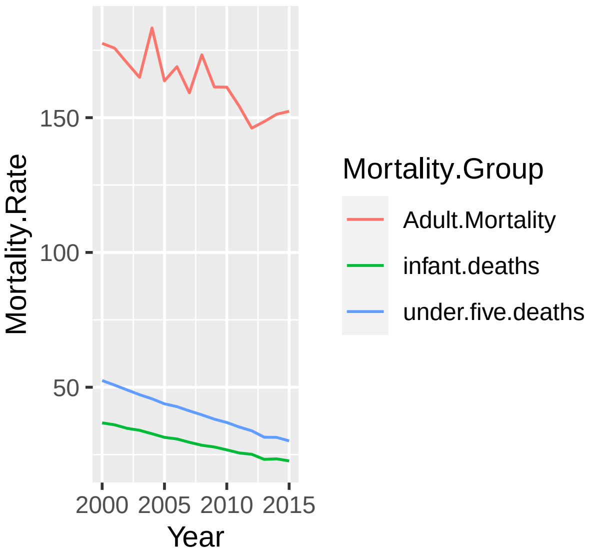
- The general life expectancy has been steadily increasing duration the year
- Average Life expectancy increase from about 67 to 71.5 in 15 years.



Finding :

- Life expectancy of Developed countries are significantly higher than that of Developing countries.





Findings :

- The mortality rate of all three age groups are generally decreasing as a whole
- The mortality rate of the adult group, however, have fluctuation within the period

`#{r} #library(Hmisc) #hist.data.frame(df) #`

Findings : - As we see, the response variable Life Expectancy is normally distributed and so our first try is to see if MLR is able to predict well. - Also, using this model and running a BIC on it, we can understand the columns that are important.

Predictor Space

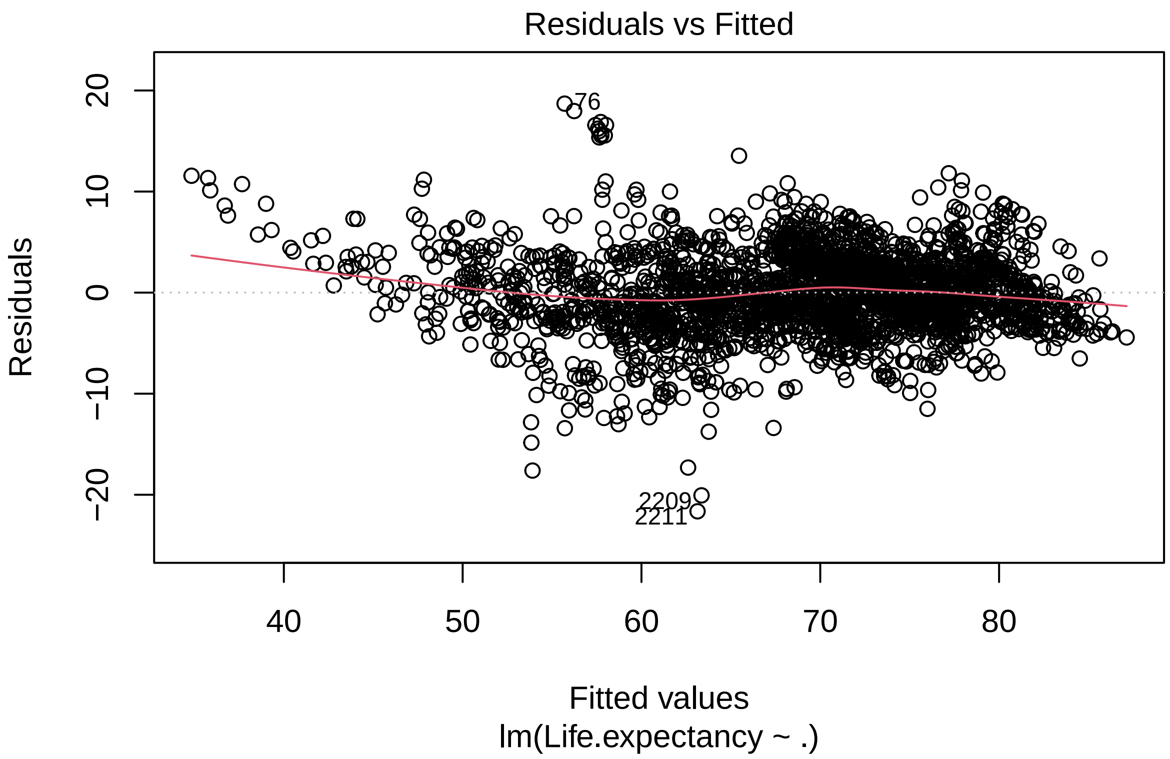
We now turn our focus to look at the different predictor variables. Following shows the correlation of different variables and their spread in the dataset.

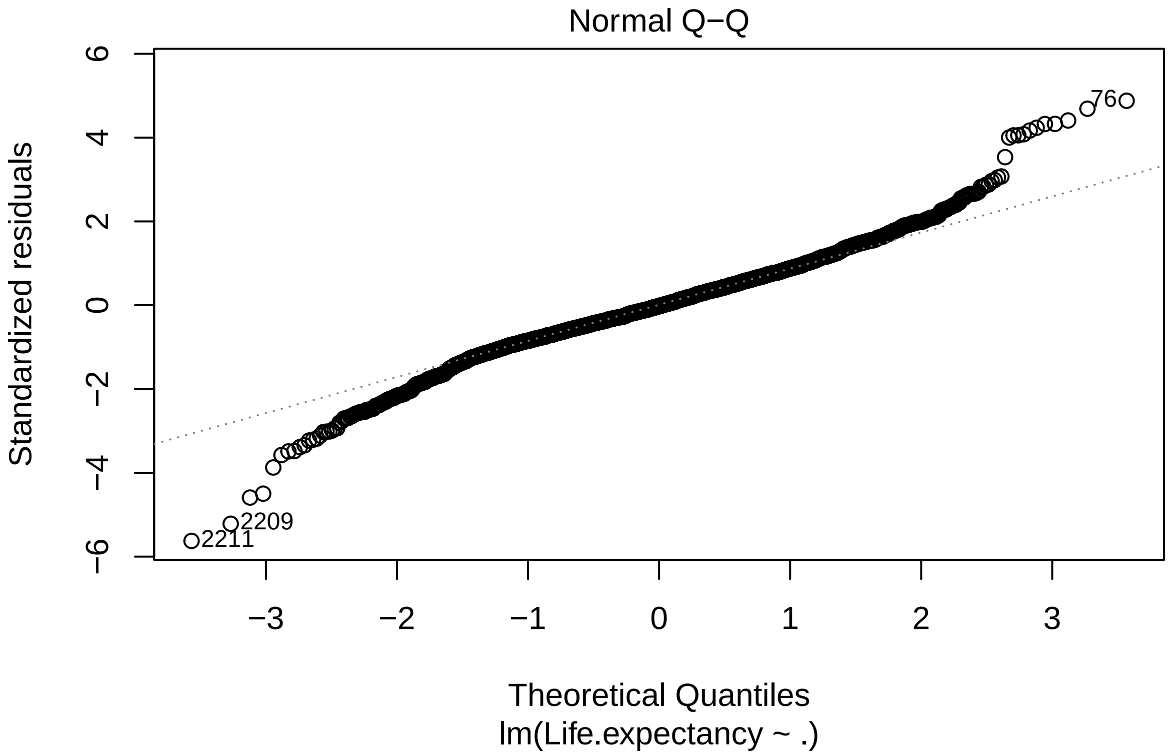
infant.deaths and under.five.deaths is nearly 100% correlated with each other and Polio is highly correlated with Diphtheria, Income composition of resources is highly correlated with Schooling. Other variables are low to moderately correlated. The response variable Life expectancy is highly correlated with Income composition, Schooling and adult mortality variables.

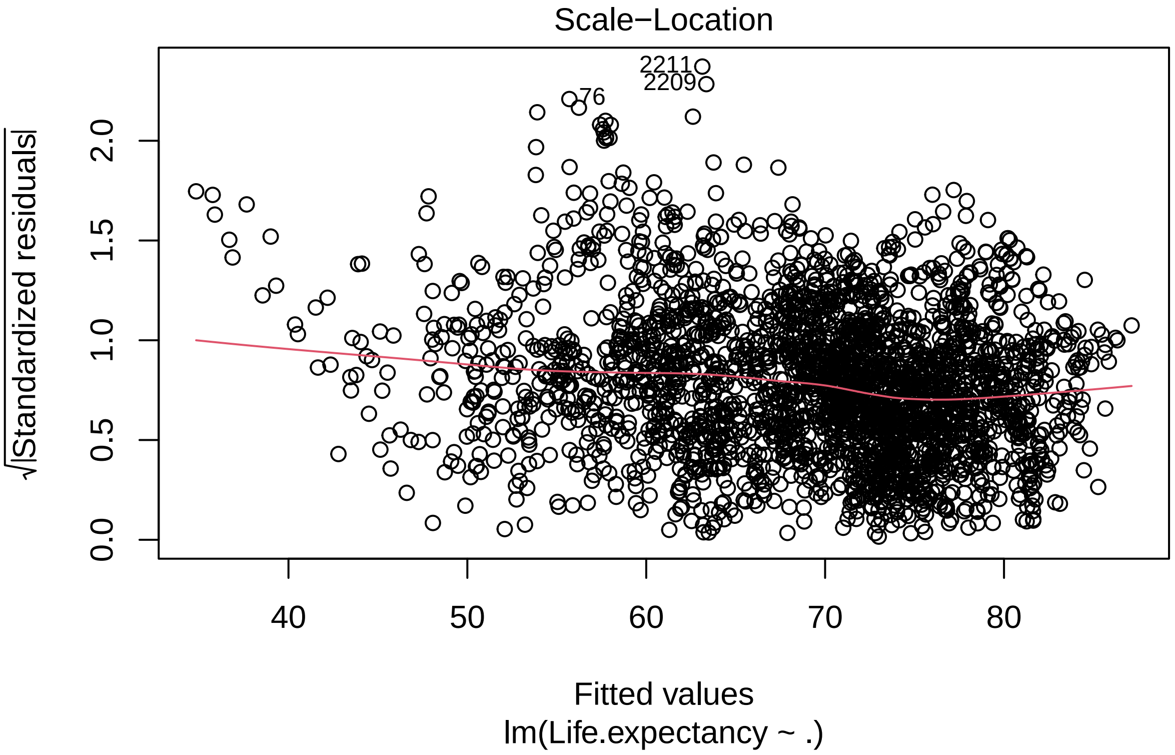
Initial Modelling and Variable Importance:

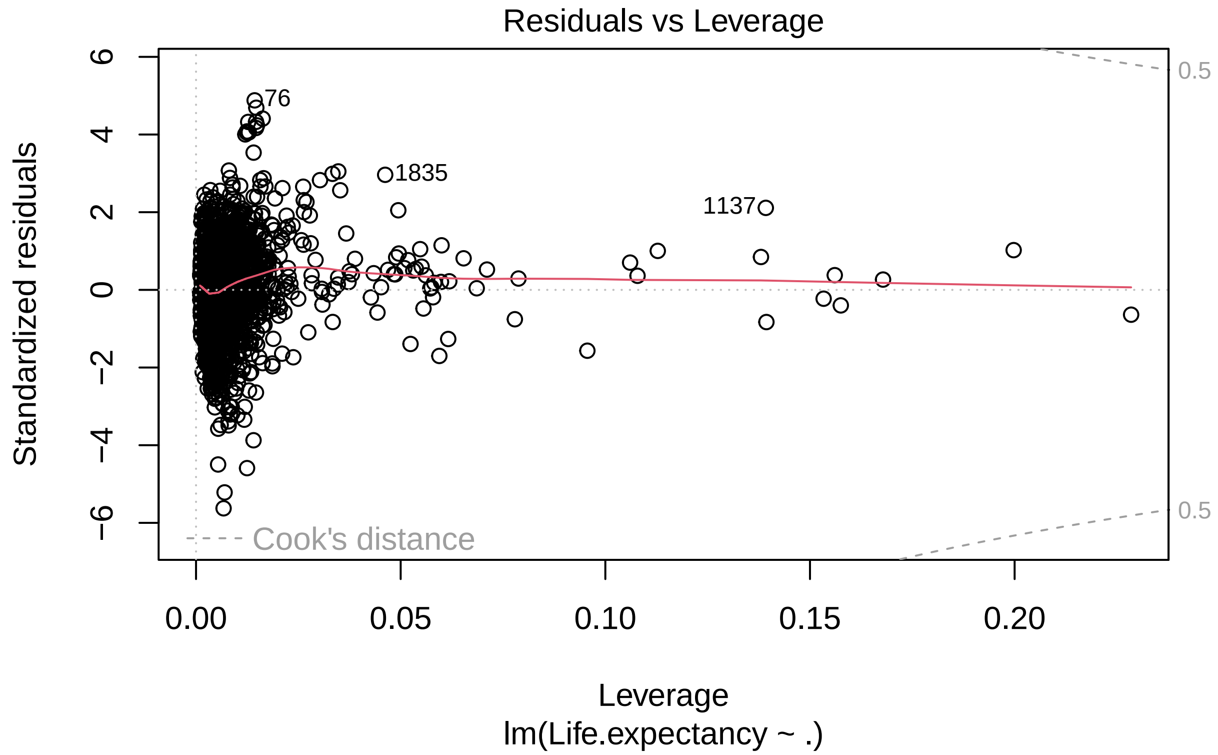
As response variable Life.Expectancy is approximately normally distributed, first step is to try lm model for this data and also run BIC to get the variable selection from the dataset.

The dataset has 20+ predictors and based on correlation plot there are correlation between the variables, BIC would help to eliminate some of the predictors that are conveying same signal as others and also explains less variability in life expectancy.


Residuals vs Fitted
Residuals
20
10
0
−10
−20
76
2209
2211
40
50
60
70
80
Fitted values
lm(Life.expectancy ~ .)


Normal Q−Q
Standardized residuals
6
4
2
0
−2
−4
−6
76
2209
2211
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Life.expectancy ~ .)





Residuals versus Fits plot: - The residuals versus fits plot would provide us with information on the residual against the fitted values in regression analysis. This could be used to identify the patterns in the residuals that may indicate the model is not capturing the relationship between our predictor and the outcome variable, therefore, allowing us to detect any non-linearity, unequal error variances and outliers. In general, we would want to see our residual randomly scattered around 0 since this indicates that the model assumption is met and is a good fit for the data. However, from the above-plotted residuals versus fits plot, we could see there is a curvature shape to our residuals and there is a presence of outliers and high leverage points on the left-hand side of the residuals versus fits the plot, This could be problematic since outliers and leverage points could have a significant impact on the regression coefficient. And the curved shape indicates that our model may be misspecified and further investigation is needed.

QQ plot: QQ-plot (Quantile-quantile plot) allow us to investigate the univariate normality of the dataset. If the points on the QQ-plot fall approximately along a straight line, it suggests that the sample comes from a population with similar distribution to the theoretical distribution that we are comparing to. From the QQ-plot that we have plotted above, the point deviates from a straight line, this indicates that our residual distribution is different from our theoretical distribution, where we could identify some outliers on both ends of our QQ-plot.

##Vigi################################

Standardized Residuals versus Fits plot:

Standardized Residuals versus Leverage plot:

Since the original model contains 20 dependent variables, this could be a concern since all variables might not be relevant to our response variable and could decrease the precision and increase the complexity/interpretability of the statistical model. Therefore we will perform BIC (Bayesian Information Criterion) variable selection to identify the most relevant variables in the data set. This method could allow us to come up with the simplest model that is still able to explain the data well and performing BIC variable selection could also help us in avoid overfitting, which occurs when a model fits the noise in data rather than the underlying relationships.

Initially, all 20 variables were used in our model and achieved an AIC score of 7642.14:

Life.expectancy ~ Year + Status + Adult.Mortality + infant.deaths + Alcohol + Hepatitis.B + Measles + BMI + under.five.deaths +Polio + Total.expenditure + Diphtheria + HIV.AIDS + thinness..1.19.years + thinness.5.9.years + Income.composition.of.resources + Schooling + Status.val + GDP_scaled + Population_scaled

After performing BIC backward step model selection method:

Life.expectancy ~ Status + Adult.Mortality + infant.deaths + Hepatitis.B + BMI + under.five.deaths + Polio + Diphtheria + HIV.AIDS + thinness..1.19.years + Income.composition.of.resources + Schooling + GDP_scaled

The BIC backward step model selection method has reduced our model's independent variable to 13 and achieved a lower AIC score of 7604.34. Since a lower AIC score signifies the regression is a better fit to the data, meaning that after removing some irrelevant variable in our data set, the simple model is still able to explain the data well well and has improved the fitting from the initial model. Also the final reduced model has an Adjusted R-squared score of 0.8296 compared to the original model's 0.8299 isn't much of a drop in the Adjusted R-squared score meaning even tho we are using fewer variables in our model, our BIC reduced model was still able to have the same amount of variability being explained by our original model.

Now going back to variable selection, this time we will be using VIF(Variance Inflation Factor) to investigate whether it is possible to come up with a better model from BIC reduced model by eliminating some highly correlated variables in the data.

- VIF(Variance Inflation Factor) is a variable selection method that is used to identify and eliminate highly correlated variables in a regression model. If the VIF value for a variable is high, it indicates that the variable is highly correlated with another predictor within the model. From the above output, we could see that variables "infant.deaths" and "under.five.deaths" are highly correlated. Since Multicollinearity can make it difficult to interpret the coefficients of the model and can reduce the overall accuracy of the model. Therefore we will have to remove either "infant.deaths" or "under.five.deaths" to resolve the multicollinearity within our data inorder to improve our model accuracy and interpretability.

Upon removing the variable 'under.five.deaths' due to the multicollinearity concern that we have seen from the VIF(Variance Inflation Factor) output is giving a lower model performance, so we will retain it moving forward. Here we update our data to our BIC-selected variable.

After doing variable reduction, checking for any clustering effects in data.

Looks like there are some clusters in the data, it could be it is because of variable “Status”’s developed vs developing. Acknowledging this information which may be helpful in future phases of model building and fine tuning. For example, if MLR would be the final model, building interaction with cluster variable and rest of data would further improve model performance.

Using our reduced model, we feed that data to Linear Model and achieve an Adjusted R-squared of 0.8296 and all variables have a p-value of less the 0.05, meaning that all of our independent variables within the model is statically significant to our dependent variable, In other words, there is strong evidence against the null hypothesis, suggesting that the observed relation between other dependent variable and independent variable is significant and real, not just due to random variation or chance.

Key Findings

Summarized below are some key findings from EDA. - Response variable is looking to be normally distributed and initial model score is ~82% which means this model is able to explain 82% of variation in Life expectancy. Its possible to use multiple linear regression for this data. From the diagnostic plots it may be seen that there is skewness in the data. - There are some variables that suffer multi collinearity (from VIF) scores. - Not all predictors are necessary to describe response variable. Model selection will be helpful.

Due to the spread of data (clustering, non-linearity of predictors w response variable, skewness in data), it is necessary to explore other models specifically non-parametric regression models.

Questions and Next Steps

1. Does the variables selected using BIC and linear model able to explain Life Expectancy adequately? A hypotheses test is required for this.
2. Is the response variable normally distributed? Shapiro-Wilk test for normality will need to be conducted for this.
3. Is Multi linear regression the best model or go with other non parametric models? From initial feedback, there are hypotheses tests available to validate this.

Need to explore further on these questions from proposal stage and conclude. 4. Understanding impact of individually controlled factors - The dataset has all predicting variables divided into 4 groups: Immunization related factors, Mortality factors, Economical factors and Social factors. Some of these factors are controllable by individuals like immunization, alcohol etc. Some of these factors are noncontrollable and macro elements like GDP. If an individual within a country want to improve life expectancy, how much is controllable/can be influenced personally? What proportion of variation in life expectancy can be explained by these variables? For example, What is the effect of “Alcohol/BMI” on the life expectancy?

5. Understanding impact of Government/Public controlled factors - From Government perspective, how are the preventive measures influencing life expectancy? What proportion of variation in life expectancy can be explained by these variables? For example, Does Higher health expenditure (column H) on Health improve life expectancy?

Extra - add if there is page available

###############PCA(suggest a simpler scoring system):##########################
##

###FA:(not working)#################
##

df2 = subset(le_dropped, select = -c(Country,GDP_scaled,Population_scaled)) head(df2)

head(df) fa_model <- factanal(df2, factors = 13)